云 南 植 物 研 究 2008, 30 (5): 597～602
Acta Botanica Yunnanica DOI: 10.3724 SP.J.1143.2008.07283

药用植物资源的整合信息数据标准规范探讨*

李丽玲, 王雨华**

(中国科学院昆明植物研究所, 云南 昆明 650204)

摘要: 从药用植物资源可持续利用对科学数据的需求出发, 分析了药用植物资源信息数据的现状, 阐明了在科研信息化环境下建立具有参考型数据库性质的药用植物资源整合信息数据库的必要性, 探索了药用植物资源数据的规范与框架, 开发了面向药用植物资源可持续利用的信息需求和共享交换的数据元标准, 为建设共享、集成的药用植物信息系统提供了统一的数据接口, 以期促进药用植物资源共享集成信息平台的建设。

关键词: 药用植物; 数据标准; 数据元; 数据共享; 参考型数据库

中图分类号: Q 948 **文献标识码**: A **文章编号**: 0253-2700 (2008) 05-597-06

Study on the Data Standard of the Integrative Information of Medicinal Plant Resources

LI Li-Ling, WANG Yu-Hua**

(*Kunming Institute of Botany, Chinese Academy of Science*, Kunming 650204, China)

Abstract: In human's research history, the scientific discoveries rely heavily on the accumulation of knowledge, information and data. China is promoting the innovative research approach and calls for the development of self-independent intellectual property on drug research and industry. However, the current status of knowledge and data on medicinal plant resources are scattered, isolated and unshared which hampers the efficient research. Based on the needs for sustainable use of medicinal plants, this study analyzed the current status of data on medicinal plants and the necessity for data synthesizing and standardizing, developed the framework and provided the standards schema for data exchange under the environment of e-Science. It is intended to construct a platform for information synthesizing and sharing.

Key words: Medicinal plant; Data standard; Data element; Data share; Reference data collection

人类科学研究史表明, 科学规律的发现依赖于大量科学数据的积累 (中国科学院科学数据库专家委员会, 2007)。在加快建设国家知识创新体系的今天, 自主知识产权的新药创制越来越离不开巨大的科学数据的积累。新药创制的重要途径之一便是从药用植物中寻找活性物质与先导化合物, 这个途径依赖于药用植物科学数据的大量积累与高效利用。但是, 目前药用植物资源的信息却呈现孤岛状态, 信息分散、积累不够、重复建设现象严重, 很难形成系统的海量数据, 阻碍着科研人员高效地获取和分析有效信息。因此在科研信息化环境下探讨具有参考型数据库性质的药用植物资源整合信息数据库的建设具有重要的现实意义。

1 药用植物资源及其信息数据的现状

药用植物是指凡能治疗、预防疾病和对人体有保健功能的植物 (中国药材公司, 1995)。不

---

* 基金项目: 云南省科技计划项目 (2000C0022Q; 2004C0021R); 国家科技基础条件平台项目 (2005DKA21006), 中科院知识创新工程信息化专项 (INF105-SDB-1-39)

** 通讯作者: Author for correspondence

收稿日期: 2007-12-04, 2008-02-15 接受发表

作者简介: 李丽玲 (1982-) 女, 在读硕士研究生, 主要从事植物资源与生物信息系统的研究。

管是在传统医药体系中，还是在现代西方医药体系中，药用植物都扮演着重要的角色。我国的传统医药主要包括传统中药、民间药（草药）以及民族药，统称为广义的中药，皆以植物、动物、矿物为药物原料，其中尤以植物应用得最多，故历史上称药物学为“本草学”(国家中医药管理局，1998)，足见药用植物在传统医药体系中的重要作用。现代西方医药体系中，药物主要是人工合成的化学药品和来自自然界的天然药物，从植物中寻找活性物质、提取有效成分一直是西药制备的一个有效途径（Balunas and Kinghorn，2005；Cragg and Newman，2003；Taylor and Berridge，2006)，即使在合成药物占主导的二十世纪中期也一直在进行（Taylor and Berridge，2006)。

在回归自然的呼声日高的今天，药用植物的研究日益受到重视（Taylor and Berridge，2006)，通过本草古籍文献（Sehgal等，1994)、民族植物学（裴盛基，1988；Patwardhan，2005）调查等途径，从传统药用植物中寻找具有活性的药物、进行新药开发已经成为新药开发的重要途径之一，也使通过古籍文献记载、民间草医传承、民间流传等方式保存的药用植物传统知识受到空前的关注（Balunas and Kinghorn，2005；Gurib-Fakim，2006；裴盛基，2007)。

在医药行业的自我发展之外，国家政策也对药用植物的研究提供了新的动力。《国家中长期科学和技术发展规划纲要（2006～2020年》把“中医药传承与创新发展”列为优先发展的主题之一，把重大疾病防治水平的提高、新药创制的突破进展等作为科技发展的重要目标之一，药用植物的可持续利用与发展迎来了良好的机遇。

与药用植物的重要性不相称的是它的资源现状和信息提供现状。

在全球生态环境破坏与生物多样性降低的背景下，药用植物资源也面临着同样的生态压力，同时由于药用植物资源与人类的密切相关性，许多野生种类已经走向濒危甚至已经灭绝，迫切需要进行保护与采取可持续的利用方式。

现代西方医药体系的建立在很大程度上改变了人类传统的药物使用方式，许多世代相传的民族、民间药用植物传统知识日渐失传，迫切需要对民族药、民间药进行系统的整理、保存，以利于保护传统知识，促进新药创制。

就药用植物的信息数据而言，有着数千年积累下来的大量的本草著作和现代研究产生的更为多样化、丰富的数据，信息量十分庞大。为了保存传统知识和有效利用信息，信息学家和学科领域专家做出了大量的努力，如对古籍文献进行电子化，建立药用植物专业数据库等。但是在这庞大的数据量面前，科研人员依然感到数据积累不够，信息贫乏。

数据积累不够的原因是多方面的。对药用植物而言，数据的量的积累相对丰富，关键在于“质”的不足，主要是信息分散、无序、缺乏共享与交换导致量的单纯积累无法满足广泛的科研需求。药用植物的相关数据主要分布于本草古籍、各类志书、现代文献以及数据库中。古籍及一些现代的本草著作及志书，其撰写方式的典型特征是“苞综诸经”(叶显纯，2004)，也就是综合多种本草古籍，在有选择地广收前人著作“精华”的基础上加上自己的见解而成书，是一个顺承的过程。各类著作的内容除了各有偏重之外，又相互引用，交错复杂，并且无法藉由数字化来解决内容的复杂性问题，本草著作中的数据利用方式始终无法摆脱“大海捞针”的困窘。在传统的书籍之外，已建立了大量的数据库，但这些数据库往往是分立的，存储的信息往往过于单一，或者信息量过少，或者数据来源不明确，无法满足全面性、权威性、易用性等要求。研究者使用现有数据库时，一般需要从数个不同的库中检索不同方面的内容，同时进行文献检索，多方寻找、多方比较，其所付出的代价并没有因为数据库的应用而降低。

2 建立药用植物资源整合信息共享的重要意义

为了解决数据整合、共享与交换问题，科学家们进行了理论与实践多方面的探索。2005年9月，美国NSB（National Science Board）发布的关于科学数据库的研究报告（NSB，2005)，将科学数据库分为研究型数据库（Research Dataset)、资源型数据库（Resource Dataset）和参考型数据库(Reference Dataset)。研究型数据库是指某一个或几个固定的项目研究产生的数据集，其中的数据

只经过有限的处理与管理，只服务于特定的研究群体，数据库获得的资助资金较少，资助周期也较短，因而没有严格遵循相关数据标准，数据规模与覆盖范围有限，数据可靠性较差。资源型数据库服务于单一的科学与工程组织或机构，其经费直接来源于相关机构，其数据通常遵循一定的数据规范。参考型数据库遵循成熟的数据标准，经费预算通常很大，多由一个或多个机构提供长期的经费支持，数据规模庞大和影响面广，服务于大范围的科学与教育机构，其典型特征是具有大范围的、多样化的用户群体，包括来自不同地域、学科与机构的科学家、学生及教育工作者。

目前，国内现有的药用植物资源数据库基本上都是研究型数据库，只有少部分的资源型数据库与参考型数据库。研究型数据库往往结构不同，数据规范不统一甚至阙如，面向的用户群狭窄并且各不相同，加上数据拥有权的限制，使信息资源的分布往往呈现“孤岛”状态，不同信息系统的数据无法共享与交换，甚至同一信息系统内部的数据也无法交换，导致信息数据的重复建设与极大浪费。打破信息孤岛限制，将研究型数据库整合成能够被广大的科研人员及教育工作者共享及重用的资源型数据库或参考型数据库，是药用植物资源科学数据库发展的趋势。

另外，整合科学数据资源也是药用植物资源数据内在关系特性的要求。世界是作为一个相互联系、相互影响的整体而存在的，科学研究中划分的不同学科只是人类在不同的认识水平、不同时空尺度上对世界的探索。随着人类对世界的认识的加深，不同学科的内在联系日益显现，必然趋向于交叉融合，对科学数据的需求也趋向于多样化、整体化。打破学科限制，整合不同学科的数据，使其能够在数据海洋中被标识、查找，满足不同的使用目的，是信息工作者以及学科专家共同面对的问题。

药用植物的数据资源整合必须建立在统一的数据标准之上。国家科学数据共享工程指出，科技数据共享是国家科技基础条件平台的重要环节和内容之一，而科学数据共享标准化工作又是实现我国科技数据共享的基本前提和保证。建立和完善科学数据共享标准体系对于国家科技基础条件平台的建立和我国经济社会的发展都极其重要(科学数据共享工程办公室，2003)。

对药用植物数据本身而言，同样具有这样的发展趋势。同时，在对数据资源赋予“药用”这一特定意义时，要加快药用植物科学数据的积累以促进新药创制，必须制定药用植物科学数据的共享规范，建设参考型数据库，在此基础上实现数据的有序积累。

3 药用植物整合信息共享的规范框架

根据科学共享工程标准体系的框架（科学数据共享工程办公室，2004），为了进行药用植物科学数据的共享与交换，其数据规范主要包括以下几个方面的内容：

1) 元数据内容：在科学数据共享核心元数据的基础上，根据药用植物研究领域自身特点，在内容信息、数据表现信息等元数据子集中扩展，为药用植物数据集的描述、标识、定位、查询、维护等提供基本的信息。

2) 数据分类与编码：按照科学数据分类与编码的一般方法与相关标准，对药用植物数据进行分类与编码，提供对数据进行无歧义理解的科学分类基础。

3) 数据模式：按照数据模式描述规则和方法，对药用植物数据的概念、组成、结构、相互关系等进行描述，在概念和逻辑层次上建立明确、统一的共享数据模式，为实施物理层次上数据共享的有效、合理的简化接口提供基础。

4) 数据交换格式：为了实现双方或多方之间交换和共享科学数据，将特定的数据元按照一定的语法组合成结构化的数据编排格式，以便能在不同数据处理系统间进行数据的无损转换。

5) 数据元目录：按照数据元标准化的原则与方法，对药用植物科学研究所涉及的数据项进行采集、分析、归类、协调等工作，对每项数据元的名称、定义、数据类型、相互关系等属性进行规范化处理，提供对药用植物科学数据的语义正确、独立、无歧义的理解。

6) 数据图示表达规范：利用一定的图示表达语言，对药用植物数据中可能使用的图示表达的内容进行表达方式的规范，如数据表达成什么样的统计图、数据表或矢量图，图线的宽度、颜色、透明性等各种属性。

共享规范各个方面的内容相辅相成，为药用植物参考型数据库系统的构建提供了从数据内容、模式到分发、共享所必需的一系列规范化的支持。其中，数据元规范是对数据进行无歧义理解的基础，是数据语义上的规范，第 4 部分对此进行探讨。

4　药用植物数据元规范

4.1　药用植物数据元的概念

数据元（data element）是指用一组属性描述其定义、标识和允许值的数据单元，在一定语境下，通常用于构建一个语义正确、独立且无歧义的特定概念语义的信息单元（中国标准研究中心，2002）。数据元由对象类、特性与表示三部分组成，是自然界中存在的对象类的特性的表示。

药用植物数据元是指药用植物科学范畴内的数据单元，对这些数据单元进行定义、标识和允许值等方面的描述，构成语义正确、无歧义的信息单元。

药用植物数据元标准为包含药用植物科学范畴内不同方面的数据、结构不同的数据库信息系统间的数据存取、共享与交换提供共同语言及和编程接口，其应用和推广有利于沟通信息孤岛，按照统一的规范，在现有的研究型数据库的基础上，整合建设资源型数据库及参考型数据库，为最大限度地实现不同系统的药用植物信息互联及共享利用奠定基础。

4.2　药用植物数据分析

药用植物数据的内容标准，特别是数据元标准，是数据共享体系的重要组成部分。确保数据在语义层上的协调一致，是整合药用植物信息建设的基础。制定药用数据元规范，必须对药用植物资源的信息需求进行分析，探求数据的范畴、形式和需求状况，在此基础上按规范化的方式进行描述，形成对数据的无歧义理解。

本文通过专家座谈会讨论、文献调研以及专家访谈等方式，揭示出药用植物信息资源作为研究对象，构成了一个完整的、立体化的知识体(图 1)，具有明确的本体范畴，可以进行共享概念模型的形式化规范描述。知识体的数据来源于药用植物研究相关的各个学科，具有纵向与横向的信息范畴及自身独特的性质，是科学研究中重要的参考体系。

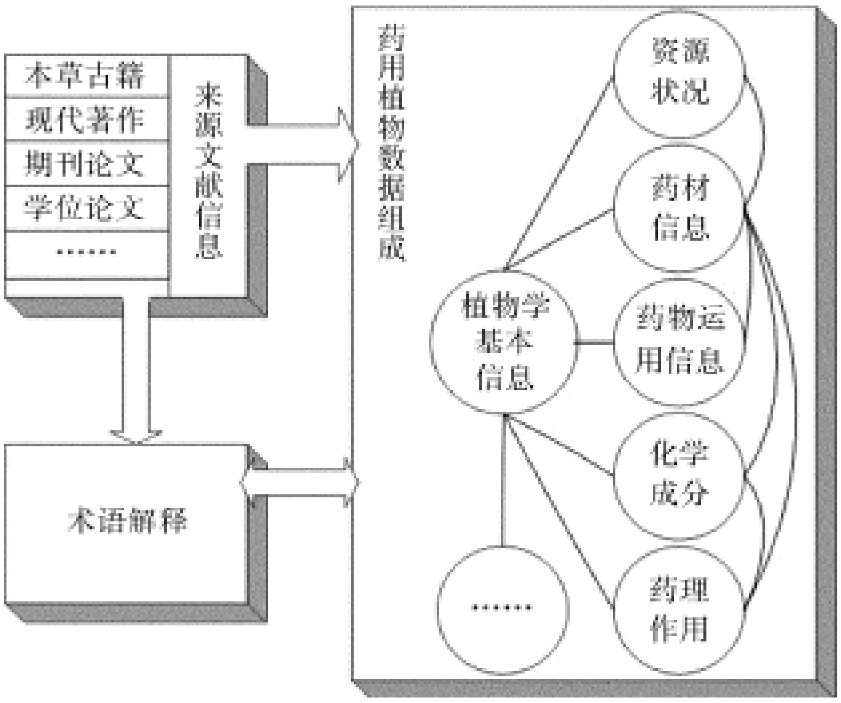

图 1　药用植物知识体结构

Fig. 1　Knowledge ontology of medicinal plant

横向信息范畴，是指药用植物不同方面的数据，主要包括如下几个方面：

1) 植物学基本数据：亦即植物物种数据，包括植物所属的科、属、种，植物拉丁名与中文名；标本信息；形态描述等，主要来源于各类植物志以及新种发表文献，是考证药用植物基源、辨识药物真伪的重要基础。

2) 资源数据：包括药用植物的地理分布、资源量、规范化种植、市场贸易等信息，是开发利用药用植物的重要参考指标。

3) 药用数据：不同的人群把植物作为药物利用时产生的相关信息数据，如药用部位、采收季节、加工方法、功能主治、药物配伍等，来源于本草学文献、民族及民间医药文献，特别是民族植物学家、民族药物学家的调查结果，是保存和利用民间流传的传统药物知识的重要资料。

4) 理化数据：物理、化学等学科手段应用到医药领域而产生的信息数据。物理数据主要用于药品质量评价，包括相对密度、馏程、熔点、凝点等；化学数据主要指药用植物所含化学成分的数据，包括成分、化学结构、分子式等。理化数据整合自现有的研究型数据库，是探究药物作用机理、进行新药创制必不可少的数据。

5) 药理作用。对药物作用及其机制的记录，数据整合自现有的研究型数据库，以及药典、研

究论文等。

纵向信息范畴，是指从信息的时域性、区域性和权威性出发，都可以把信息分成不同的层次，也就是纵向的关系。

信息的时域性是指信息产生的年代特征。对药用植物而言，可以把其信息数据大致划分为古代信息、现代信息。古代信息指来自古籍本草的信息。古籍主要指 1912 年以前在中国书写或印刷的、具有中国古典装订形式的书籍（国家技术监督局，1987）。现代信息则相对于古籍本草中的信息而言。

信息的区域性由药用植物分布的区域性决定，如果是广域分布、广域利用的，那么它的相关信息将分布在不同区域，其多样性也分布在不同区域；如果是狭域的，那么它的信息将局限分布在某一区域，其独特性也更明显。

信息的权威性也指信息的可靠性，由信息的初始来源、验证过程等决定其层次。若以等级形式表出，可以粗略分为以下几级：

1) 国家级的标准化的信息，以法定形式颁布或者得到同行一致认定。如由国家相关部门组织编写，并以法定形式颁布施行的《中华人民共和国药典》，又如《中国民族药志》、《中国植物志》等。

2) 一是由国家组织的全国性资源普查所得的资料的整理出版物，如 20 世纪 70 年代的《全国中草药汇编》、八九十年代的《中国中药资源丛书》等，为大量学科专家进行资源普查、初步试验之后，协同整理而成的著作，具有一次科学文献的特性。二是由研究者广搜博采，多方考证而成，且被后世公认为权威性著作的本草学著作，如《本草纲目》、《滇南本草》等。

3) 地方级的标准化的信息，或地方性的具有重大学术意义的书籍，如《云南植物志》、《中国彝族药学》等，起到对层次 1 的补充作用。各类学术期刊上发表的一次文献中的数据，是科学工作者的工作结晶，具有重要的学术意义。

4.3　药用植物数据元标准化

按照数据元标准化的原则，主要使用自下而上的数据元提取法提取药用植物知识体中的数据元素进行规范化描述。药用植物数据集的数据元包括基于药用植物关系型参考数据库的实体对象、属性、联系等元项。基于“数据元是自然界中存在的对象类的特性的表示”这种认识，首先识别出对象类，然后寻找对象类的特性，再把特性表示出来。

在药用植物信息知识体中，根据面向需求的药用植物数据分析的结果，结合建立数据模式的实践，识别出 6 个相互关联的主要对象类，即植物物种信息类、药材信息类、化合物信息类、药理信息类、资源信息类、方剂信息类；2 个辅助对象类，即参考文献类和术语类。在每一个对象类下，根据实际情况划分子类，如方剂信息类下又分为传统中药、民族药和民间药方剂信息子类。

对象类的特性相当于“实体 - 关系”模型中的“属性”。如对象类“方剂”的特性“功效”、“配伍及用量”等，对象类“药材”的特性“药用部位”、“性”、“味”等。

为了在数据元集里唯一标识每个元素，采用 9 位的内部标识符（编号），见图 2。编号的每级分类顺序号从 01 开始，无某级分类的，则该级编号为 00，该级分类名称为空，数据元顺序号从 001 开始编号，实体对象类别用两位大写字母标识。

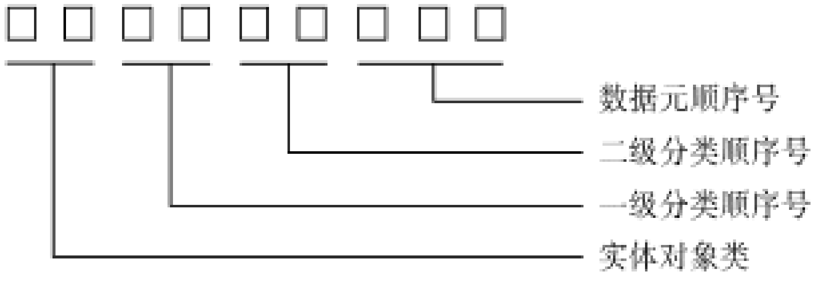

图 2　药用植物数据元编号

Fig. 2　Code of medicinal plant data element

药用植物的每个数据元由一组属性来描述，这些属性是描述数据元的“元数据”（数据的数据），见表 1，其中“M”指“必选”，“O”指“可选”，“C”指“条件选”。

对药用植物数据元的属性定义，如果有相关国家标准、行业标准的，则尽量采用标准格式与表达，否则多方查询辞书、志书、专业著作、文献等资料，结合常用表达习惯而定。对数据元的注册与维护，由传统的注册机构运行改为开放式运行机制，拟在 Internet 公开接受修改与维护。

表1 数据元属性词典

Table 1 Dictionary of data element attribute

类别	属性	约束	类别	属性	约束
标识类	标识符	C	定义类	定义	M
	名称	M	表示类	表示类别	M
	英文名称	M		表示形式	M
	版本号	C		数据类型	M
	同义名称	O		最大长度	M
	相关环境	C		最小长度	M
关系类	分类模式	O		表示格式	C
	关键字	O		值域	M
	相关数据参照	O		语言	O
	关系类型	C		备注	O

5 结语

本文从药用植物资源的可持续利用的信息需求出发，探讨了药用植物资源信息共享与交换以及科学数据积累的方法与途径，按照相关方法制定药用植物的数据元规范，提供了药用植物信息共享与交换的数据接口，是构建药用植物参考型数据库的基础性的第一步，为药用植物资源的整合信息共享打下重要的基础。

美国国家科学基金会（NSF）已经明确提出：未来，美国在科学和工程上的国际领先地位将越来越取决于其在数字化科学数据的优势上，取决于通过成熟的数据挖掘、集成、分析与可视化工具将其转化为信息和知识的能力（NSB, 2005)。当前，世界各国都认识到科学数据不仅是科研活动的产物，而且是支撑国家科技创新的重要资产（中国科学院科学数据库专家委员会，2007)。在信息化的时代，建立药用植物资源整合信息可以为药物筛选提供方向，有利于充分发挥传统优势对现代药物创新的作用，为新药创制提供资源信息保障，是建设创新型国家的重要基础。为了完善新药创制的数据支撑体系，还需在统一的科学数据共享规范之下，以新药创制技术为指导，整合细胞水平和分子水平的数据以及实验数据，为提高自主创新能力提供完整而有力的科学数据支撑体系。

〔参考文献〕

中国药材公司，1995．中国中药资源丛书——中国中药资源［M］．北京：科学出版社

中国标准研究中心，2002．信息技术 数据元的规范与标准化 第1部分：数据元的规范与标准化框架，GB T 18391［R］．1-2002 ISO IEC 11179-1：1999

中国科学院科学数据库专家委员会，2007．中国科学院科学数据库资源整合与持续发展研究报告［Z］

叶显纯，2004．《神农本草经》初探［J］．中医文献杂志，**22**（2）：1—2

国家技术监督局，1987．古籍著录规则．GB T 3792．7-1987［R］．北京：中国标准出版社

国家中医药管理局，1998．中华本草［M］．上海：上海科学技术出版社

科学数据共享工程办公室，2003．科学数据共享标准化工作指南(草案稿)．SDS T 1002-2004 科学数据共享工程技术标准［R］

科学数据共享工程办公室，2004．标准体系及参考模型［R］．SDS T 1001．1-2004 科学数据共享工程技术标准

Balunas MJ，Kinghorn AD，2005．Drug discovery from medicinal plants［J］．*Life Sci*，**78**（5）：431

Cragg GM，Newman DJ，2003．Plants as a source of anti-cancer and anti-HIV agents［J］．*Ann Appl Biol*，**143**：127—133

Gurib-Fakim A，2006．Medicinal plants：Traditions of yesterday and drugs of tomorrow［J］．*Mol Aspects Med*，**27**（1）：1—93

NSB，2005．Long-Lived Digital Data Collections：Enabling Research and Education in the 21st Century（ID：CSD4071）［Z］．Contributed by Organizations or Campuses（2005）

Patwardhan B，2005．Ethnopharmacology and drug discovery Perspectives of ethnopharmacology［J］．*J Ethnopharmacol*，**100**（1-2）：50—52

Pei SJ（裴盛基），1988．Ethnobotany and the exploitation of plant resources［J］．*Acta Bot Yunnan*（云南植物研究），**Suppl I**：135—144

Pei SJ（裴盛基），2007．Brief disscussion on ethno-medicine research and new-drug development of China［J］．*J Yunnan Coll Tradit Chin Med*（云南中医学院学报），**30**（3）：1—7；**30**（4）：4—7

Sehgal R，Ackers JP，Warhurst DC，1994．Looking for drugs in ancient texts［J］．*Nature*，**372**（6502）：124

Taylor S，Berridge V，2006．Medicinal plants and malaria：an historical case study of research at the London School of Hygiene and Tropical Medicine in the twentieth century［J］．*Transactions of the Royal Society of Tropical Medicine and Hygiene*，**100**：707—714